ITS 2.0 Metadata and Machine Translation

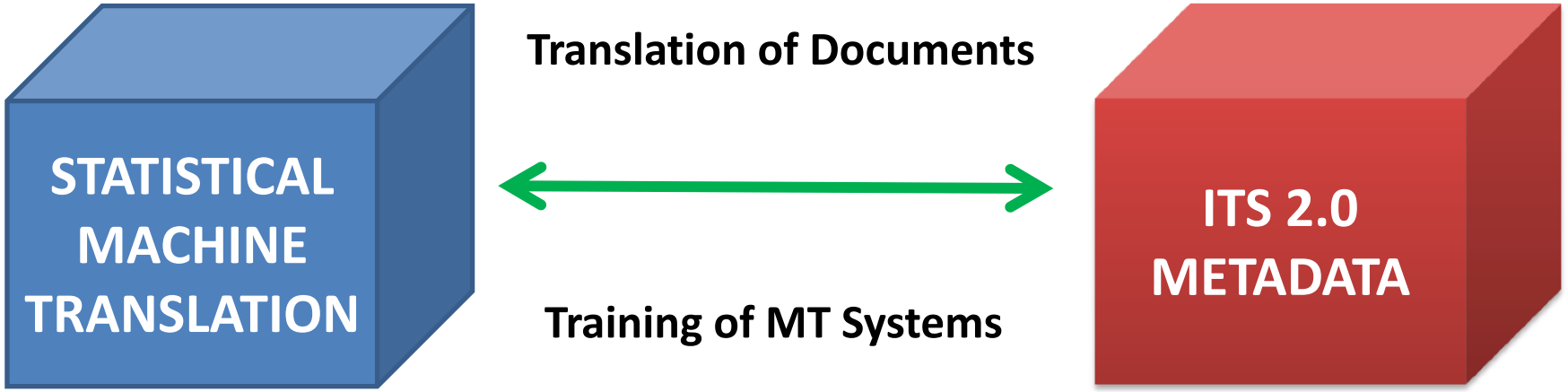
Ankit Srivastava & Declan Groves

Centre for Next Generation Localisation,

School of Computing, Dublin City University

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Overview



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

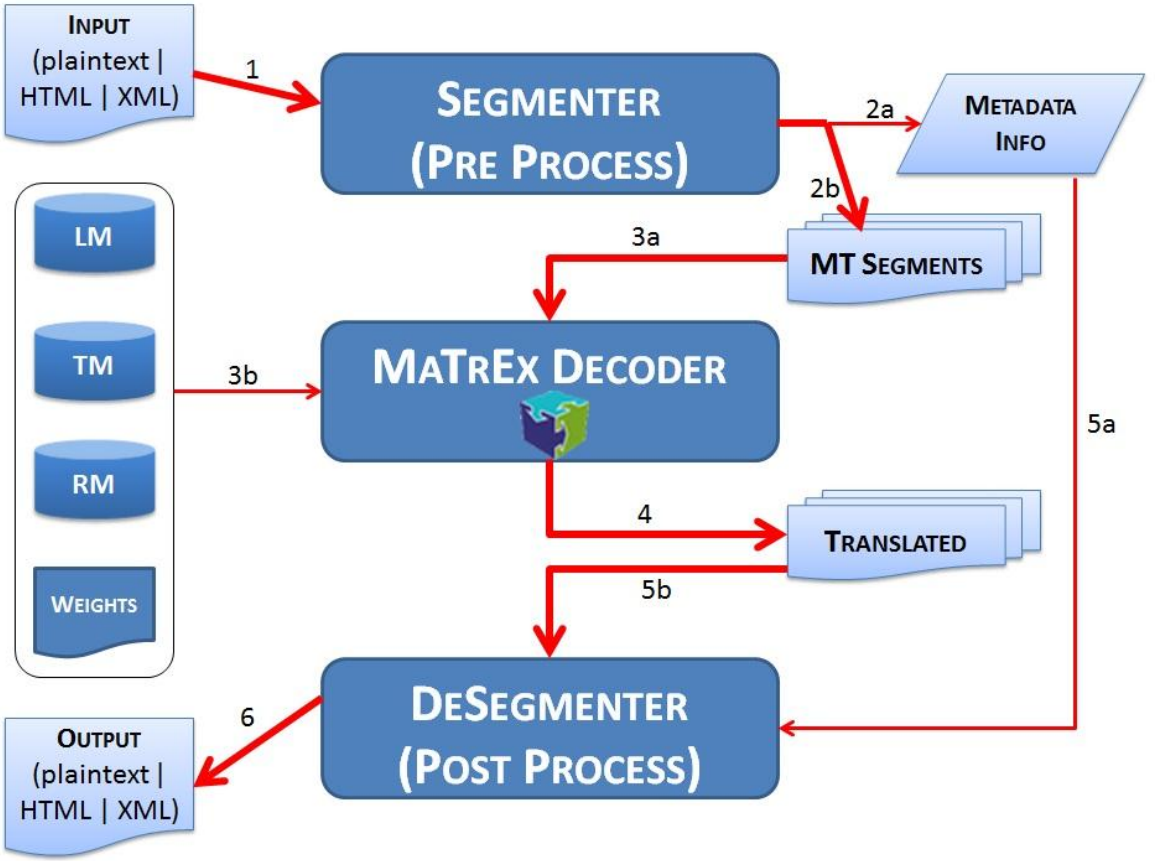
A Statistical Machine Translation (SMT) System developed in-house at DCU using the open-source Moses decoder

Segmenter takes as input ITS 2.0 tagged document, parses it, and generates segments to be translated [Pre Process]

Decoder translates the segments

Desegmenter rebuilds the document by merging the translated segments with relevant ITS 2.0 metadata [Post Process]

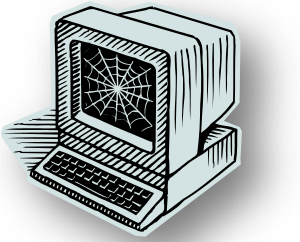
A set of pre-processing and post-processing scripts wrapped around the MaTrEx Decoder enables us to translate ITS 2.0-tagged documents in HTML / XML / XLIFF



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Technical Demo

- System developed as part of WP4 (Online MT Systems)
  - with Linguaserve and Lucy
- Poster 6 at Multilingual Web Workshop – Rome
  - **Simple Segment Machine Translation**
    - Receives input from TCD's CMS-LION [XLIFF]
    - Online Demo at http://www.cngl.ie/mlwlt/



ITS 2.0 Data Category	Usage / Benefits to the System
Domain	Enables seamless application of domain-tuned MT engines
Language Info	Enables easy identification of source / target language
Locale Filter	Enables locale-specific translation operations
MT Confidence	Enables scores to be displayed accurately and automatically to PEs
Provenance	Enables localisation workflow managers to compare performance
Terminology	Enables terms to be identified for translations to be enforced in MT
Translate	Ensures identification of text fragments that need to be preserved

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

DEMO

Play Video

Link At:

http://www.computing.dcu.ie/~asrivastava/docs/WebService_testrun.wmv

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

ITS 2.0 Metadata in MT

MT DECODING

- Demonstrated successfully the translating of documents tagged with ITS 2.0
- Usefulness in MT
  - Translate *forced translations*
  - Domain *model adaptation*
  - MT Confidence *useful for PE*
- Translation Web Service

MT TRAINING

- Can documents tagged with ITS 2.0 be leveraged to train MT engines & components?
- Training Web Service

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

ITS 2.0-aware MT Training (1/2)

- ITS 2.0 bilingual data is collected via Cocomore's CMS and passed to DCU's MaTrEx MT System
- Content
  - Text in both source language as well as target language
  - Tagged with specific translate, domain tags, etc.
  - Segmented at sentence / paragraph / document level
- Process
  - Retrain models by processing out ITS 2.0 tags (**v1**)
    - cf. Panacea (EU FP7) Web Services for MT Component Training
  - Retrain models by using translate / domain ITS2.0 tags for domain-tuned / NER-sensitive TMs (**v2**)

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

ITS 2.0-aware MT Training (2/2)

- ITS 2.0 Data Categories Used
  - Translate, Terminology, Language Information, Domain, Provenance
- Benefits
  - SMT Systems will now be trained on data which are ITS 2.0 tagged instead of mere plain text
  - The ITS 2.0 markup provides key information to drive the reliable extraction of domain-specific content from both XML and HTML5.
  - MT Systems trained on domain-specific data allows for a potentially more accurate translation

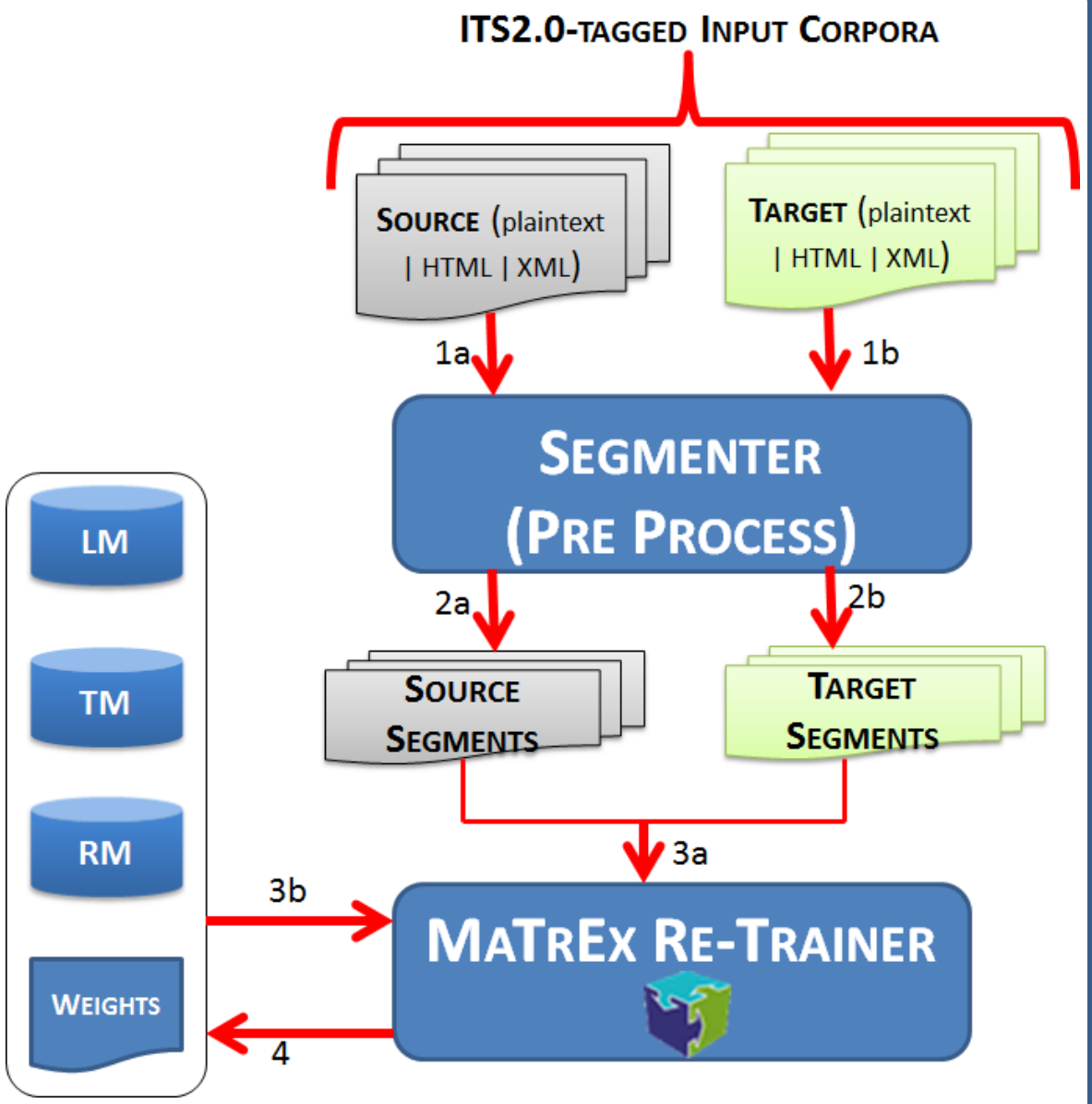
The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

MATREX

SMT components that could be potentially trained / retrained include a Language Model (LM), Translation Model (TM), Reordering Model (RM), and associated feature weights

Segmenter takes as input ITS 2.0 tagged documents (sentence-aligned parallel content), parses it, and generates bilingual segments for training [Pre Process]

Retrainer processes bilingual segments to augment pre-existing SMT Models



The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Work in Progress / Future Work

- MaTrEx Translation Service
  - Fine-tuning / Snags [*mid April 2013)*
- MaTrEx Retrainer Version 1 *[end* March 2013*]*
  - **Passive** Use of ITS 2.0
  - ITS 2.0 tags processed before training
  - Demonstration on “toy data”: Fr-En, Es-En
- MaTrEx Retrainer Version 2 *[end September 2013]*
  - **Active** Use of ITS 2.0
  - ITS 2.0 tags domain and translate used in TMs
  - Tuning MT systems with ITS2.0 parallel data from Cocor (De-Fr) generated in task 5.1

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Thank You!

asrivastava@computing.dcu.ie

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Supplementary Information

EXTRA SLIDES

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Panacea (FP7)
Soaplab Web Services at DCU

moses_build_phrase_table (WSDL) *Builds a phrase table using moses (takes plain or factored models input files).*

Run service

Inputs

source_corpus as URL / direct data or local file — Browse...

source_language

target_corpus as URL / direct data or local file — Browse...

target_language

Report

http://srv-cngl.computing.dcu.ie/panacea-soaplab2-axis/

Workplan

Work package number [53]	WP5	Type of activity [54]	SUPP
Work package title	Deep Web Information and MT Training		
Start month	4		
End month	21		
Lead beneficiary number [55]	3		

Delive-rable Number [61]	Deliverable Title	Lead benefi-ciary number	Estimated indicative person-months	Nature [62]	Dissemi-nation level [63]	Delivery date [64]
D5.1.1	Drupal MT Training Module	10	8.00	P	PU	15
D5.1.2	XLIFF Deep Web MT Training Exporter	9	5.50	P	PU	15
D5.2	Metadata-Aware MT Training Tools	3	4.75	P	PU	15

- V1 of D5.2 to be produced in M15
  - In-line with DoW
- V2 of D5.2 to be produced in M21
  - End of WP according to DoW

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

14

Task 5.2 Breakdown

S. No.	Task / Activity	F	M	A	M	J	J	A	S	O	N	D
D5.2	Metadata-Aware MT Training	■	■	■	■	■	■	■	■			
D5.2a	MaTrEx retrainer v1	■										
1	Investigate feasibility of Soaplab	■										
2	Customize Pre-Process for Corpus	■	■									
3	Test-Run Trainer on Toy Data		■									
4	EN-FR Retraining		■	■								
5	EN-ES Retraining		■	■								
6	Cocomore Data Retraining			■	■	■	■					
D5.2b	Retraining Module v2						■	■	■			

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.